SAND REPORT
SAND2003-2812
Unlimited Release
Printed August 2003

Genomes to Life Project Quarterly Report February 2003

Grant S. Heffelfinger

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2003-2812
Unlimited Release
Printed August 2003

Genomes to Life Project
Quarterly Report
February 2003

Grant S. Heffelfinger
Materials and Process Sciences Center
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0887

Abstract

This SAND report provides the technical progress for the first quarter (through February 2003) of the Sandia-led project, "Carbon Sequestration in Synechococcus Sp.: From Mol.ecular Machines to Hierarchical Modeling," funded by the DOE Office of Science Genomes to Life Program.

Understanding, predicting, and perhaps manipulating carbon fixation in the oceans has long been a major focus of biological oceanography and has more recently been of interest to a broader audience of scientists and policy makers. It is clear that the oceanic sinks and sources of CO_2 are important terms in the global environmental response to anthropogenic atmospheric inputs of CO_2 and that oceanic microorganisms play a key role in this response. However, the relationship between this global phenomenon and the biochemical mechanisms of carbon fixation in these microorganisms is poorly understood. In this project, we will investigate the carbon sequestration behavior of *Synechococcus* Sp., an abundant marine cyanobacteria known to be important to environmental responses to carbon dioxide levels, through experimental and computational methods.

This project is a combined experimental and computational effort with emphasis on developing and applying new computational tools and methods. Our experimental effort will provide the biology and data to drive the computational efforts and include significant investment in developing new experimental methods for uncovering protein partners, characterizing protein complexes, identifying new binding domains. We will also develop and apply new data measurement and statistical methods for analyzing microarray experiments.

Computational tools will be essential to our efforts to discover and characterize the function of the molecular machines of *Synechococcus*. To this end, molecular simulation methods will be coupled with knowledge discovery from diverse biological data sets for high-throughput discovery and characterization of protein-protein complexes. In addition, we will develop a set of novel capabilities for inference of regulatory pathways in microbial genomes across multiple sources of information through the integration of computational and experimental technologies. These capabilities will be applied to *Synechococcus* regulatory pathways to characterize their interaction map and identify component proteins in these

pathways. We will also investigate methods for combining experimental and computational results with visualization and natural language tools to accelerate discovery of regulatory pathways.

The ultimate goal of this effort is develop and apply new experimental and computational methods needed to generate a new level of understanding of how the *Synechococcus* genome affects carbon fixation at the global scale. Anticipated experimental and computational methods will provide ever-increasing insight about the individual elements and steps in the carbon fixation process, however relating an organism's genome to its cellular response in the presence of varying environments will require systems biology approaches. Thus a primary goal for this effort is to integrate the genomic data generated from experiments and lower level simulations with data from the existing body of literature into a whole cell model. We plan to accomplish this by developing and applying a set of tools for capturing the carbon fixation behavior of complex of *Synechococcus* at different levels of resolution.

Finally, the explosion of data being produced by high-throughput experiments requires data analysis and models which are more computationally complex, more heterogeneous, and require coupling to ever increasing amounts of experimentally obtained data in varying formats. These challenges are unprecedented in high performance scientific computing and necessitate the development of a companion computational infrastructure to support this effort.

More information about this project, including a copy of the original proposal, can be found at www.genomes-to-life.org

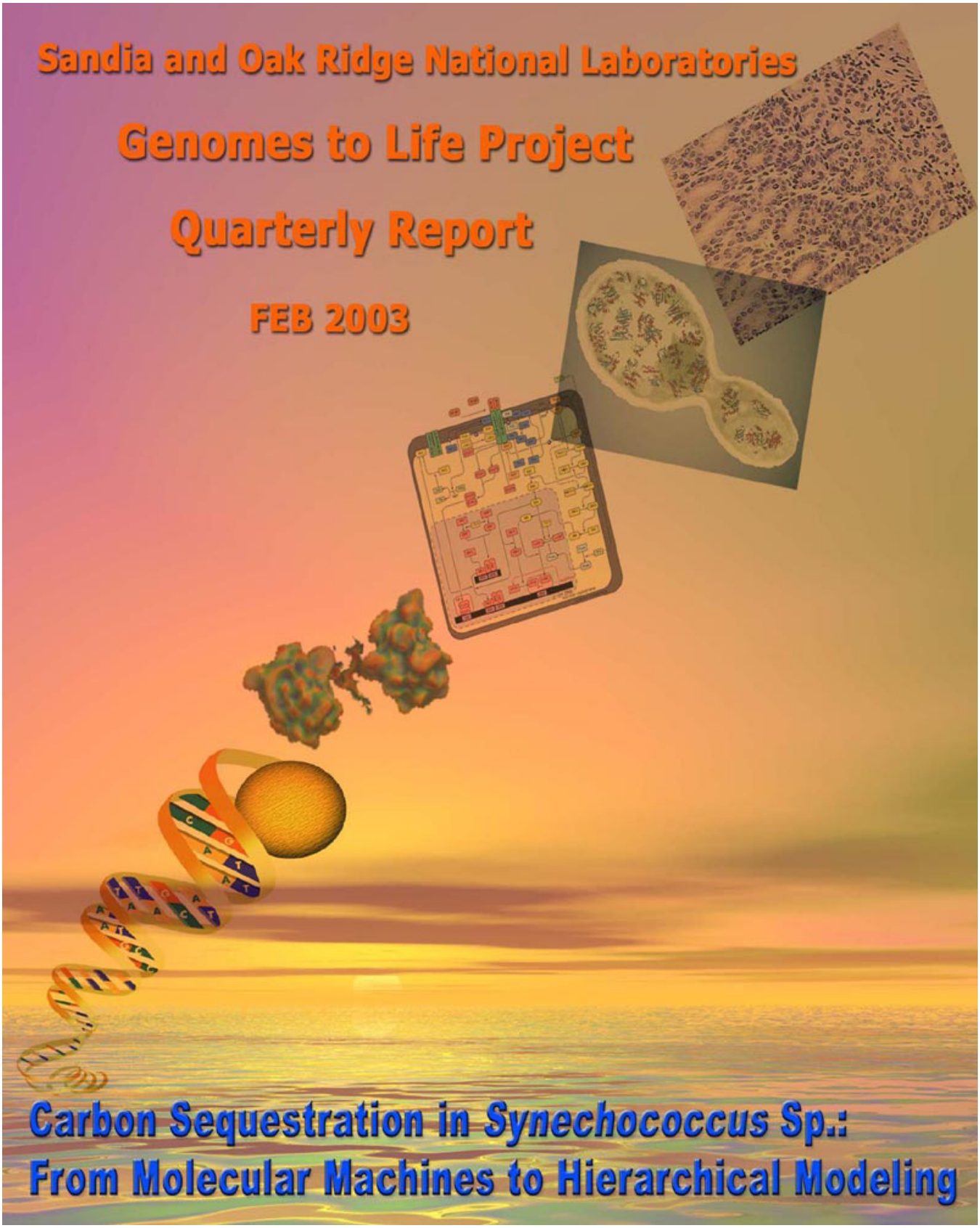
Acknowledgment

We want to gratefully acknowledge the contributions of the members of the GTL Project Team as follows:

Grant S. Heffelfinger[1]*, Anthony Martino[2], Andrey Gorin[3], Ying Xu[3], Mark D. Rintoul[1], Al Geist[3], Hashimi M. Al-Hashimi[8], George S. Davidson[1], Jean Loup Faulon[1], Laurie J. Frink[1], David M. Haaland[1], William E. Hart[1], Erik Jakobsson[7], Todd Lane[2], Ming Li[9], Phil Locascio[2], Frank Olken[4], Victor Olman[2], Brian Palenik[6], Steven J. Plimpton[1], Diana C. Roe[2], Nagiza F. Samatova[3], Manesh Shah[2], Arie Shoshoni[4], Charlie E. M. Strauss[5], Edward V.Thomas[1], Jerilyn A. Timlin[1], Dong Xu[2]

*Author to whom correspondence should be addressed (gsheffe@sandia.gov)

1. Sandia National Laboratories, Albuquerque, NM
2. Sandia National Laboratories, Livermore, CA
3. Oak Ridge National Laboratory, Oak Ridge, TN
4. Lawrence Berkeley National Laboratory, Berkeley, CA
5. Los Alamos National Laboratory, Los Alamos, NM
6. University of California, San Diego
7. University of Illionois, Urbana/Champaign
8. University of Michigan
9. University of California, Santa Barbara



Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences.

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences.

Project Management Update

Budget

As of Feb 1st, 2003, we have received just 19% of our FY02 and FY03 project budget due to the congressional budget impasse and continuing resolution and uncertainty in the DOE Office of Advanced Scientific Computing (OASCR) budget. These difficulties have significantly complicated the difficulty in placing the contracts between Sandia and our ten partner institutions as we have had very little funding to work with on a month to month basis. However, our hard work is paying off and we are emerging from those difficulties and now have funding to all of the primary participating institutions. As the rest of the FY02 and FY03 budget arrives (we are told that the OASCR budget issues have now been resolved and it appears that the congressional impasse is about to be resolved) we will move quickly to send funding to the rest of our partners. Throughout this report, the budget figures refer to what we have reserved for the listed partners, although in some cases, that funding has just arrived.

Building a Sense of Project

We have had two large group meetings since the announcement of our award, a 1-day project kickoff meeting in Santa Fe in August, 2002, hosted by the National Center for Genomic Resources, and a 2-day project meeting in October, 2002, hosted by Oak Ridge National Laboratory. Our next 2-day team meeting is scheduled in March and will be held in Salt Lake City. We have established successful monthly executive team teleconferences to discuss project management issues (primarily the budget, to date) as well as monthly technical team teleconferences. In addition, many of our participants have traveled to visit other team members at other institutions. All of these efforts have significantly helped further refine our collaboration and technical focus.

Milestones

In spite of our technical difficulties, we have made progress toward our milestones, although largely in the context of simply getting work started and establishing the collaborative activities needed to enable the project to succeed as a whole.

Report Executive Summary

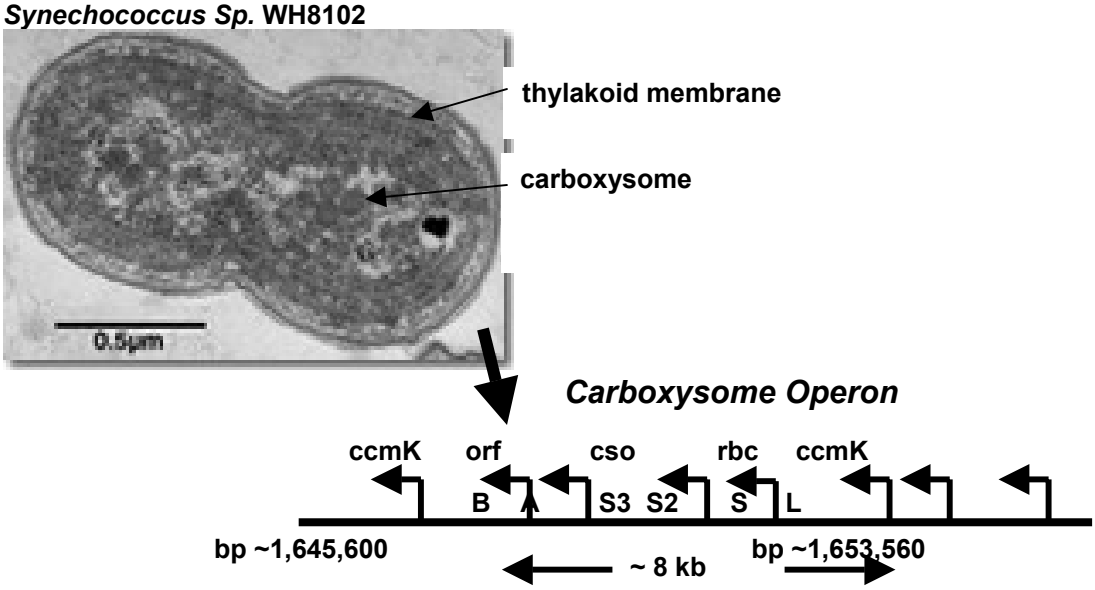
In order to accommodate the longer spin-up time required for experimental efforts, we provided a disproportionate portion of our available funding for our experimental effort. Thus this effort has made significant progress in several areas including isolating carboxysomes, developing methods to elucidate and verify protein interactions, investigating the expression of ABC transporters as a function of histidine kinase and response regulator knock-outs, and developing and applying state-of-the-art techniques for the accurate measurement and analysis of microarray data.

Our computational biology core, comprised of three subprojects focused on molecular machines, regulatory networks, and computational systems biology, has also progressed in spite of our budget difficulties. Work in this area ranges from developing new bioinformatics methods for analysis of protein-protein interactions to conducting conducted genome-scale protein structure and function predictions on all ORF's of *Synechococcus* sp. and two related genomes *Procholorococcus* MIT and MED, and carrying out a simple, trial simulation of carbon transport and fixation in a cylindrical representation of an individual *Synechococcus* via the solution of a suite of reaction-diffusion equations.

Finally, while our Computational Biology Work Environments and Infrastructure subproject has been focused on building our own project's web-based computational infrastructure (e.g. a project web site with a password-protected electronic notebook to facilitate collaboration and data sharing), we have also developed a MIAME schema for microarray data focused on the unique characteristics of the *Synechococcus* data to be provided by our experimental effort.

Subproject 1: Experimental Elucidation of Molecular Machines and Regulatory Networks in *Synechococcus* Sp.

Synechococcus Sp. **WH8102**



An electron microscope picture of the photosynthetic marine organism *synechococcus* WH8102 and a schematic gene diagram of the carboxysome operon. The thylakoid membranes and carboxysomes are both components of a complex appartus responsible for inorganic carbon fixation. More is understood of the thylakoid membranes that regulate the electron transport cycle than of the carboxysomes that participate in the carbon reduction cycle (Calvin-Benson-Bassham cycle). The schematic gene diagram of the carboxysome operon represents our current understanding.

This work was funded in part or in full by the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org).

Introduction

It is the goal of this work to elucidate the mechanisms of carbon transport to the carboxysome and carbon fixation in the carboxysome at a proteomics level based on what we know about the genomics. We hope to provide a structural and functional understanding of carboxysome multiprotein complexes. In this first quarterly report we outline our initial steps in reaching our goals.

Accomplishments

Carbon fixation in the carboxysome: Carboxysomes are polyhedral inclusion bodies that consist of a protein shell surrounding ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO). While RuBisCO regulates photosynthetic carbon reduction, the function of the carboxysome is unclear. The carboxysome may either actively promote carbon fixation by concentrating CO_2 or may passively play a role by regulating RuBisCO turnover. The presence of carbonic anhydrase, an enzyme that regulates the equilibrium between inorganic carbon species, in the carboxysome would suggest an active role in carbon concentration, but experimental results are mixed. No clear biochemical evidence for a link between carbonic anhydrase and the carboxysome exists in WH8102.

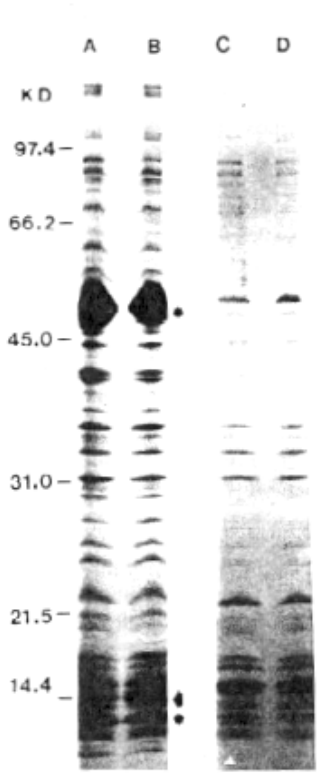


Figure 1. Carboxysome characterization of *synechococcus* PCC7942 by SDS-PAGE showing the presence of multiple polypeptides of undetermined identity. Price et al., Plant Physiol. 100, 784-793, 1992

The schematic of the carboxysome operon above is a starting point for our proteomics-based studies. The operon shows the location of the RuBisCO gene (rbc L and S) proximal to a number of genes coding for proteins found in the carboxysome shell. The list of genes may not represent all proteins that make up the carboxysome, though, as SDS-PAGE analysis suggests the presence of many more polypeptides (Figure 1). Our efforts will hopefully provide a more complete characterization of the composition of the carboxysome by purifying carboxysomes from 8102, separating the proteins by SDS-PAGE, and analyzing the bands by mass spectrometry.

This quarter we have made progress in isolating carboxysomes. We have prepared a multi-step purification protocol involving carboxysome isolation (culture and bulk up cells, followed by recovery and lyseing cells in a French press, removal of whole cells and membrane fragments, separation of carboxysomes from lysates (by high speed centrifugation) and purification (re-suspend carboxysomes in separation buffer and separate by sucrose gradation). We have also prepared techniques to separate shell proteins from RuBisCO by sonication and sucrose gradation. We have also started testing the protocols on small amounts of cells and are bulking up larger cultures needed for characterization.

In addition to elucidating carboxysome composition, we intend to characterize protein-protein interactions within the carboxysome. Our efforts include determining protein interaction partners and characterizing the protein binding motifs that mediate protein interactions. We have taken several steps to elucidate protein interactions and developed a strategy to genetically manipulate known carboxysome shell proteins with biochemical tags. We are attempting to fix HA tags to both the upstream and downstream ccmK genes and the rbc L gene separately. Genetic tags are being

prepared by PCR. They will be constructed with convenient restriction digest sites in the primers and placed in each gene as defined by the digest sites. At least two sites will be attempted for each gene. Whole operon gene segments containing the tagged genes will be shuttled into an appropriate expression system. As an alternative, we have developed gene knock-in strategies in order to make use of endogenous regulation of the operon. These efforts will proceed concurrently to the expression system attempts. Tagged proteins generated by the tagged genes will be isolated by affinity purification, and protein binding partners to the tagged proteins determined by SDS-PAGE separation and mass spectrometry. To date, the strategy is determined, the primers ordered, and the molecular biology is progressing. Additionally, protocols for plating 8102 for transformation are being tested.

It is essential to verify protein-protein interactions by using complimentary techniques. In addition to immunoprecipitation techniques, we have started work on developing bacterial 2-hybrid systems (cousin to the more familiar yeast 2-hybrid systems). Here, we will isolate each gene in the carboxysome operon and transfer them into the bacterial 2-hybrid system. Primers for isolating the genes and the molecular biology are in progress.

Transporters and Transportation of Inorganic Substrates: ABC transporters are a family of proteins that carry out influx and efflux of inorganic substrates such as nutrients in the cell. The process is effected by environmental stimuli detected by the histidine kinase, response regulator signal transduction system. The signal transduction system, then, may regulate ABC transporters. As a part of the Microbial Cell Project and now Genomes to Life, we are attempting to understand the expression of ABC transporters as a function of histidine kinase and response regulator knock-outs. We will began by measuring gene expression profiles by gene microarray techniques.

This quarter we have begun work in producing viable knock-out cultures. Additionally, we have contracted out for the construction of a whole genome microarray chip. Work is in progress.

Our project is attempting to develop state-of-the-art techniques that include accurate measurement and analysis of microarray data. We are developing hyperspectral scanning and multivariate curve resolution algorithms. This quarter, we have used the equipment budget to purchase some of the hardware needed for the construction of the GTL dedicated hyperspectral microarray scanner. We have continued to make improvements in the computational speed of the multivariate curve resolution algorithms allowing faster extraction of spectral species in the hyperspectral image data. Through our use of yeast genome microarrays we have gained a deeper knowledge of several anomalies that trouble the microarray community, such as black holes and dyes separation. This information will be used in the following stages of the GTL microarray studies to avoid these anomalies when possible and correct the data when not. Lastly, we have submitted an example for our microarray efforts of the Minimum Information About a Microarray Experiment (MIAME) protocol. This protocol is becoming a scientific standard required for publications.

Progress Towards Milestones

In reference to the milestones submitted in the original proposal, this quarter we have completed or are working on the following milestones:

(1) *Synechococcus* cultures are established.
(2) Tagging carboxysome proteins is in progress.
(3) Synthesis and calibration of microarray chips is in progress.

Collaboration With Others

A number of external collaborations have been established this quarter. We have begun a scientific interaction with Dean Price and Murray Badger of the Australian National University. Professors Price and Badger are preeminent researchers in the area of cyanobacteria inorganic carbon transportation. We have established a materials exchange program that will allow us to do cross organism comparisons. Additionally, we are acting as partner investigators on a research grant submitted by Professor Badger to the Australian Research Council entitled, 'The Structure and Function of Cyanobacterial Carboxysome Multiprotein Complexes and the Role in Carbon Sequestration in Cyanobacterium.' This partnership will allow us to participate in a student exchange program. Finally, we have begun experimental planning with Grant Jensen at Cal Tech to do electron microscopy and cryoelectron tomography studies of the carboxysome. Professor Jensen's work has been featured by DOE as Figure 1.8 of the Report on the Imaging Workshop for the Genomes to Life Program.

We have been able to solidify strong interactions between project members this quarter. Interactions between the experimental biologists, data analysis team, and the information systems team has produced a first draft of the MIAME protocol for microarray information dissemination. Additionally, culturing and plating techniques have been exchanged, microarray experiments coordinated, and a number of meetings for scientific information exchange completed. Examples of these meeting include but are not limited to discussions regarding bacterial 2-hybrid strategies, inorganic carbon transportation in aquaporins versus transporters for modeling applications, and synthesis of RuBisCO for NMR studies.

Publications and Presentations

(1) Heffelfinger, G.S., Martino, A., Gorin, A., Xu Y., Rintoul M.D. et al., Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling, OMICS 6(4), 305-330, (2002).

(2) Martinez, M. J., Aragon, A. D., Rodriguez, A. L., Weber, J. M., Timlin, J. A., Sinclair, M. B., Haaland, D. M., Werner-Washburne, M., "Identification and removal of contaminating fluorescence from commercial and in-house printed DNA microarrays," Nucleic Acids Research, 31, 4, in press, 2003.

(3) Timlin, J. A., Haaland, D. M., Sinclair, M. B.,Van Benthem, M. H, Martinez, M. J., Werner-Washburne, Margaret, "Hyperspectral Imaging and Multivariate Analysis of Gene Expression Data," 14th International Genome Sequencing and Analysis Conference, Boston, MA, October 2-5, 2002.

(4) Haaland, D. M., Timlin, J. A., Sinclair, M. B.,Van Benthem, M. H., "Multivariate Curve Resolution for Hyperspectral Image Analysis: Applications to Microarray Scanner Fluorescence Images," Chemometrics in Analytical Chemistry (CAC2002), Seattle, WA, September 22-26, 2002.

(5) Haaland, D. M., Timlin, J. A., Sinclair, M. B.,Van Benthem, M. H., "Hyperspectral Fluorescence Image Scanning of Microarrays for Improved Gene Expression Analyses," Federation of Analytical Chemistry and Spectroscopy Societies (FACSS), Providence RI, October 10-13, 2002.

(6) Haaland, D. M., Timlin, J. A., Sinclair, M. B.,Van Benthem, M. H, "Improving Microarray Analysis and Discovering Sources of Artifacts with New Hyperspectral Microarray Scanner," NIH invited presentation, Bethesda, MD, December 11, 2002.

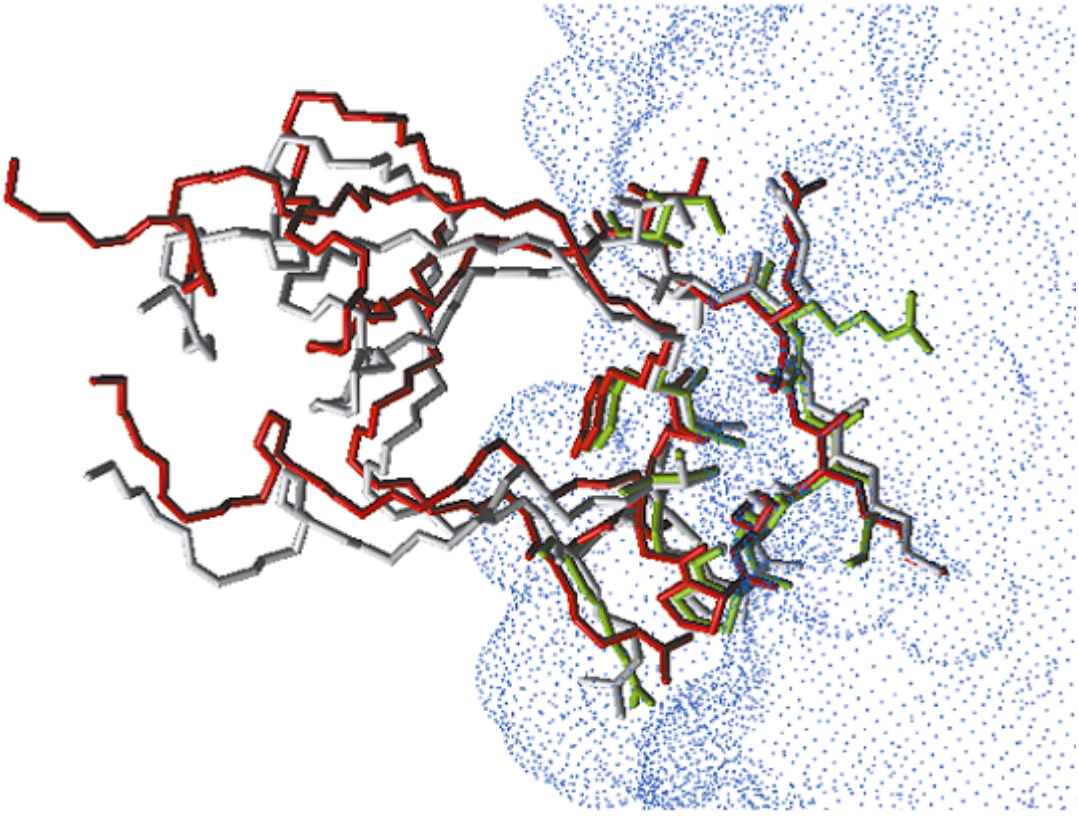
(7) Haaland, D. M., Timlin, J. A., Sinclair, M. B.,Van Benthem, M. H, Martinez, M. J., Aragon, A. D., Werner-Washburne, M. "Multivariate Curve Resolution for Hyperspectral Image Analysis: Applications to Microarray Technology," Proceedings of the SPIE, Spectral Imaging: Instrumentation, Applications, and Analysis, Vol 4959, 2003.

Budget Report (8-02 thru 1-03)

The money allocated to date has been:
SNL: $51K plus $285K in FY02
ORNL: $20K
UCSD/Scripps: Contract is being negotiated
Michigan: Contract is being negotiated

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences.

Subproject 2: Computational Discovery and Functional Characterization of *Synechococcus* Sp. Molecular Machines



This work was funded in part or in full by the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org).

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences.

Introduction

This effort is an investigation of the protein-protein interactions on several levels and from several research directions, involving sequence analysis approaches, computational interfaces to direct experimental methods, fast prediction of protein folds and detailed atom-atom simulation. The questions asked at different levels are intricately interconnected and, correspondingly, the perused approaches not only complement each other, but also interact and enhance alternative research methods. During this quarter we have laid out strategies, draw main blueprints of the ideas to research and implement, and actually started implementation of new tools for protein functional characterization and computational analysis of mass-spectroscopy data.

Our work attracted substantial interest during Super Computing 2002 in Baltimore. Two posters highlighted the Subproject: the overall concept of the molecular machine computational discovery was presented in Molecular Machines of Life: Discovery of Protein-Protein Complexes through High Performance Computing (Andrey Gorin) and the working prototype for protein domain functional characterization – *SVMMER* **-** was presented by Nagiza Samatova.

Accomplishments

Analysis of Protein-Protein Interactions by Bioinformatics Methods

In the area of bioinformatics methods we have focused on the following development goals: (1) taxonomy of high-resolution protein function groups; (2) system for inferring protein functions from genomic sequence. We have tested the performance of the system using various statistical cross validation techniques and in the next quarter will apply it to the prediction of functional roles for uncharacterized Synechococcus sp ORFs.

A new method was devised for hierarchical multi-resolution functional classification of protein sequences. The entire protein space was partitioned into taxonomy of protein functional groups. The taxonomy was created automatically with manual curation using proteins with known functions from the Swiss-Prot. The highly conserved (high resolution) protein functional groups are at the leaves of the taxonomy. The method is based on extracting discriminating features that are specific to each node of the hierarchy and applying a machine learning method to classify its children nodes. The best method is selected based on performance results for several discriminative methods, such as support vector machines, association rules, and decision trees. We demonstrated that this method allows us to functionally differentiate highly conserved functional groups. The method is also scalable with the number of functional groups since at each node only the locally computed features are considered. At the same time, it is the specificity of the locally extracted features that is the main driver behind the method's accuracy. Finally, we used our classification method to predict functional roles for a number of uncharacterized Synechococcus sp ORFs. The entire methodology is implemented as part of the protein function characterization system, called SVMMER.

Our accomplishments on developing computational methods for inferring protein functions provide a basis for the Task 4 of our Project 2. This methodology will be expanded in the future to handle protein complexes. So far the decision is based mainly on sequence information. Our next step in this direction is to expand its ability to handle structural and biochemical properties of proteins. Our plans for the next quarter will be focused on two areas: (1) to develop methods for extracting statistically significant features for the purpose of functional characterization and protein-protein interactions; (2) to develop a contingency table based approach for predicting putative pairs of interacting proteins and apply the methodology to a subset of genes in *Synechococcus* sp.

Computational Interfaces for Direct Experimental Methods

Direct experimental data are indispensable resource for any work on protein molecular machines. Mass spectroscopy is one of a very few approaches that hold great promise to provide high flow of such data. With this understanding we are concentrating our research on computational methods for mass spectroscopy. New computational algorithms and technologies are urgently needed in this area to enable mass-spec analysis of protein-protein complexes, especially if we aim to high throughput characterization of peptides in complex protein mixtures. The central problem is the heavy dependence of the existing computational algorithms (SEQUEST, MASCOT) on database lookups for potential peptide candidates. The database must encompass the whole bacterial genome in question, and the procedure moves through three stages: (1) the candidate peptide are selected by the overall mass of the experimental sample; (2) theoretical MS spectrum for each candidate is calculated; and (3) the theoretical spectra compared with the experimental one using scoring functions borrowing ideas from image comparison.

The traditional approach runs into insurmountable difficulties when the experimental sample has two peptides connected by a chemical linker. By extending the existing approach, the search for initial candidates (by overall mass) would require to scan through N*N peptide combinations for the database containing N peptide fragments. Taking into account that the typical N value for bacterial genomes will be in millions, we arrive to the set of $\sim 10^{12}$ potential peptide candidates, with the corresponding grow of complexity in two other stages of the traditional technology. There is also proportionally larger negative contribution from the usual deficiencies, steaming from database errors and posttranslational protein modifications.

Our solution to this fundamental problem is the development of MS de novo sequencing technology, where the spectrum analysis centers on the individual spectra lines. De novo tools identify protein segments without the list of potential candidates, eliminating the need for 10^{12} database lookups. There is also a very strong potential to overcome database errors and posttranslational artifacts, if sufficiently robust and efficient algorithms will be implemented.

We put forward a set of new ideas for MS de novo sequencing algorithm design. The approach is two layered. First, we are developing a set of algorithms and software tools for extensive decoding analysis of the experimental MS spectra (running title for the package is MS-VIEW). MS-VIEW identifies all types of the present ions, collects exhaustive sets of data (precision of ion mass reproduction, relative contributions of various types, ion modifications, sequence and length dependent probabilities of the bond breaking, and so forth) and, finally, performs extensive statistical analysis. On the second stage, the learned statistical rules will be integrated into our MS-CHAIN algorithm for efficient peptide identification by de novo. In the current quarter we have designed and prototyped significant part of MS-VIEW package (~ 20 functions and tools). The performed analysis of more than 2600 experimental MS-spectra solidly confirmed our design concepts and lead to the interesting findings in the properties of tandem MS spectra. In the following quarter we plan: (1) complete MS-VIEW implementation; (2) analyze "positive" data set from several research centers and from various MS instruments; (3) develop further MS-CHAIN algorithms and start design of MS-CHAIN package.

We also develop computational solutions for NMR methods involving high throughput structural and dynamic characterization of protein-protein interactions. In particular, we will pursue: the development of automated RDC-NMR methods for high throughput assignments and characterization of relative domain alignments in protein-protein complexes. The main idea here to utilize our novel approaches developed previously for nucleic acids, surmounting many specific challenges of the protein complexes.

Our accomplishments in this area address deliverables for the Task 2, where one of the main goals is close integration of the experimental data and creation of the computational algorithms, which would allow direct recording of the protein-protein interactions in high throughput fashion.

Methods for Computational High Throughput Characterization of Protein Interactions

In addition to sequence comparison one of our main focuses is development of the structure-based methods capable either to confirm the sequence-based inferences or to model the actual binding interface. At the most specific level we would like to be able to start with unbound crystallographic structures of the proteins and produce models of the likely binding conformation. Alternatively we would like to use structure prediction methods to infer the structure of protein and then use it to create hypotheses about which proteins might interact.

The latter approach has the following steps. First, the large set of potentially interacting protein domains will be filtered to a modest size set by sequence base genomic methods described above. For any of these proteins with unknown structure we predict the structure of the domains using our Rosetta algorithm. At the second step we use a structure comparison algorithm (MAMMOTH) to search for other proteins on similar structure. This helps us assess both the functional annotation of these 'unknown' proteins as well as make structure based inferences about likely binding partners. Once we reach a level where we are sure two proteins bind we can then attempt to dock these based on structural homology models.

The first step in the described process requires the creation of structure prediction pipeline based on Rosetta and Mammoth. Such pipeline would merge domain division algorithms, secondary structure prediction algorithms, sequence comparison algorithms, fragment assembly algorithms (Rosetta), and structure comparison algorithms. Also since the prediction algorithms are not unique we also must have a database system to store and curate the accumulated inferences. Structure prediction is computationally expensive. We are currently assembling a couple hundred node super computer on which we will build this genomic scale pipeline. Ultimately we will include into our tool set also PROSPECT pipeline (developed in SP3), which is very similar in the designed functionality, but very different in the underlying algorithms, and therefore could be expected significantly complement and enhance overall tool robustness and precision. The development of the pipeline contributes toward FY03 goals of the Task 2 and 4.

All-atom Molecular Simulation of Protein Interactions

During the reporting period this part of the Subproject have been focused on 3 efforts: parallel tempering for peptide conformation, optimization algorithms for docking, and transporter characterization in *Synechococcus*.

Parallel tempering is a simulation technique that can find low-energy conformations of molecular systems with many local minima. It has had particular success at predicting structures of short peptide chains and small proteins. Multiple copies of a solvated peptide chain are simulated at different temperatures. Monte Carlo rules are used to exchange temperatures between copies, enabling the peptide to overcome energy barriers. We have added a tempering capability to our parallel molecular dynamics code LAMMPS, so that P=M*N processors can be used to simulate M copies of the peptide, with each simulation running on N processors. Thus far we have tested our new code on a peptide 5-mer, Met-enkephalin, which has been well characterized in the tempering literature. We are still working on the temperature control algorithm we use (Nose/Hoover), to insure the different copies equilibrate properly between exchanges. We also need to develop a post-processing capability based on principal component analysis (PCA) techniques to diagnose the tempering output and verify that a full conformational search has occurred. Once these enhancements are in place, we will model the small peptide chains used in phage display experiments (SP1) that bind to protein domains from *Synechococcus*.

Our docking goal is to model peptide/protein and protein/protein interactions in conjunction with the phage display experiments of SP1. These computations will be performed with our docking code PDOCK, using structure-based algorithms that dock peptides to proteins of known structure and evaluate their potential binding affinities. This quarter we have enhanced PDOCK by integrating it with the SGOpt optimization toolkit. SGOpt algorithms will enable us to sample possible docking conformations of the peptide at 3 levels: (a) prediction of flexible backbone conformations (using the tempering results as input), (b) prediction of flexible side-chain conformations, and (c) local flexible refinement of the entire peptide. We have completed the first step of the integration process, which is to have the two codes communicate and to use SGOpt optimizers to perform localized flexible optimization of the ligand. We have compared the results of this technique to the previous SIMPLEX implementation and found the new code is roughly 2x faster on our test cases. Our next task is to use the genetic-algorithm (GA) optimizer in SGOpt to sample from peptide side-chain rotomer libraries, to perform an effective conformational analysis of the side chains.

Our initial efforts at modeling membrane transport of inorganic carbon into *Synechococcus* have been to characterize what is known about such transport. Two classes of transporters (passive porin transporters and active ABC transporters) move CO_2 (carbon dioxide) and HCO_3^- (bicarbonate) into cells. Initial homology searches for such transporters in the *Synechococcus* genome have not produced any hits, despite evidence in the literature that aquaporin inhibitors decrease CO_2 uptake by *Synechococcus*. Our hope has been to find *Synechococcus* proteins homologous to porin proteins of known structure. This would enable us to build an atomic-scale model of a prototypical passive *Synechococcus* transporter and anaylyze its permeability and transport properties with our molecular modeling codes. We are discussing other possible solutions to this problem with the SP1 experimentalists and the Jakobssen's U Illinois group. We are also contacting Murray Badger and Dean Price at the Australian National University who have published extensively on carbon-concentration mechanisms in aquatic photosynthetic organisms.

Progress Towards Milestones

As we have outlined in our report in our R&D we closely follow to FY03 deliverables. Specifically, we have significant progress toward three FY03 goals: (1) develop parallel tempering technology, all-atom docking models for flexible peptide chains; (2) implement incorporation of the experimental data (NMR and mass-spectroscopy); (3) create catalog of proteins in *Synechococcus* that are relevant to specific metabolic pathways.

Collaboration with Others

Our work on the integration of the mass-spec experimental data brings close and productive collaboration with Subproject 1 (PI Tony Martino) and ORNL-PNNL GTL project Center for Molecular and Cellular Systems (PI Michelle Buchanan). Center scientists Dong Xu and Tema Fridman have significantly contributed to our development of mass-spec analysis package. We also closely work with mass-spectroscopists Gregory Hurst and Robert Hettich.

Tremendous opportunities for synergetic collaboration between Subproject 3 pipeline based on PROSPECT (PI Ying Xu) and our efforts for fast fold characterization and all-atom molecular modeling were outlined above.

Currently we develop collaboration with Ruben Abagyan's laboratory at Scripps Research Institute, with the intention to apply ICM technological platform for protein docking part of out project.

The functional characterization studies have been performed in collaboration with Computational Biology Group led by Natalia Maltsev (Argon National Laboratory) and Brian Palenik from Scripps.

Publications and Presentations

1. Nagiza F. Samatova, Gong-Xin Yu, Praveen Chandramohan, Hoony Park, George Ostrouchov, Al Geist, Natalia Maltsev. "Hierarchical feature extraction based approach to functional differentiation of highly homologous protein functional groups" (to be submitted to *Bioinformatics* journal).

2. Nagiza F. Samatova The Poster and Demo of the SVMMER have been presented at SC2002, November 16-23, 2002, Baltimore, MD.

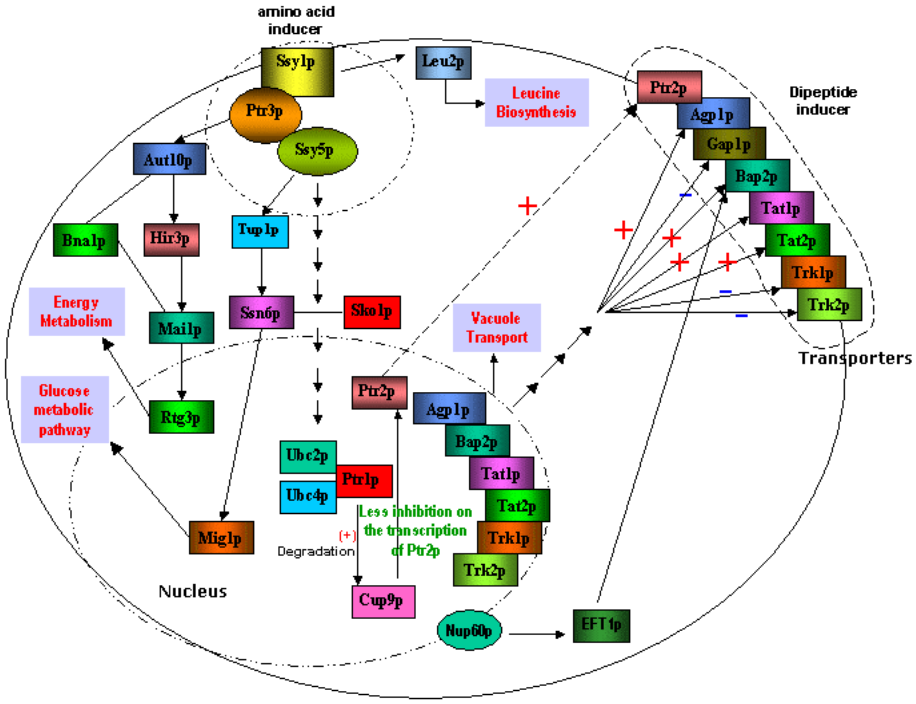
3. Andrey Gorin "Molecular Machines of Life- Discovery of Protein-Protein Complexes through High Performance Computing". Poster presentation, Super Computing 2002, Baltimore, November 16-23, 2002.

4. Andrey Gorin and Hashim M. Al-Hashimi "Combinatorial Assignment Procedure for RNA and proteins." Keystone Conference: Frontiers of Biomolecular NMR, Taos, NM, February 4-10, 2003.

Budget Report (8-02 thru 1-03)

The money allocated to date has been:
SNL: $85K
ORNL: $90K
LANL: $37K

Subproject 3: Computational Methods Towards The Genome-Scale Characterization of *Synechococcus* Sp. Regulatory



A pathway model of amino acid transporter.

This work was funded in part or in full by the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org).

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences.

Introduction

In living systems, control of biological function occurs at the cellular and molecular levels. These controls are implemented by the regulation of activities and concentrations of species taking part in biochemical reactions. The complex machinery for transmitting and implementing the regulatory signals is made of a network of interacting proteins, called regulatory networks. Characterization of these regulatory networks or pathways is essential to our understanding of biological functions at both molecular and cellular levels. Traditionally, study of regulatory pathways is done on individual basis through ad hoc approaches. With the advent of high-throughput measurement technologies, e.g., microarray chips for gene/protein expression and two-hybrid systems for protein-protein interactions, and bioinformatics, it is now feasible and essential to develop new and effective protocols for tackling the challenge of systematic characterization of regulatory pathways. The key goals of this project are to develop novel computational methods to

1. significantly improve existing computational capabilities for characterization of regulatory pathways,
2. significantly improve current capability for extracting biological information from microarray gene expression data,
3. significantly improve the existing capability for identifying co-regulated genes,
4. implement these computational capabilities on HP computers for genome-scale applications, and
5. investigate a selected set of regulatory pathways in Synechococcus through applications of our new computational tools and experiments.

Accomplishments

The following has been accomplished in the first quarter of this project:

1. conducted genome-scale protein structure and function predictions on all orfs of *Synechococcus* sp. and two related genomes *Procholorococcus* MIT and MED. All prediction results are at http://compbio.ornl.gov/PROSPECT/syn/
2. made operon structure predictions in the three genome through comparative analysis of conserved gene relationships in the three genomes and other cyanobacterial genomes.
3. produced detailed structure/function predictions, using our own prediction tools, on 300 selected orfs, which are suspected to be essential to the key biological functionality of *Synechococcus* sp genome.
4. built initial structure models of proteins involved in carboxysome complex in *Synechococcus* sp.
5. prepared and presented a set of 8 2-hr. lectures on statistical experimental design and process control for our university collaborators. The course was video taped and the tapes will be distributed to interested GTL collaborators.
6. performed a statistical analysis of nine hybridized microarrays with the 250 gene *Synechococcus* microarray to determine the repeatability of the experiments. Also aided in the specifications for the whole genome *Synechococcus* microarrays to be generated and hybridized at TIGR.
7. analyzed Affymetrix microarray data to test the methods for generating simulated microarray data with realistic errors that will be used to test and evaluate the relative performance of various bioinformatics algorithms to be applied to microarray data.

Progress Towards Milestones

Genome-scale predictions and analysis of *Synechococcu*s sp and its related genomes has generated highly useful information for our planned work to build biological pathways. Our detailed analysis on the 300 orfs allows us to begin to put together possible pathway models relevant to these genes. We are currently in the process of building a few pathway models relevant to ABC transporters in the genome, using data generated from structure/functional analysis of the orfs and information generated from mining other data

resources, including protein-protein interaction data, operon structure predictions, and microarray gene expression data.

Efforts are on-going to build effective computer tools to predict protein-protein interaction maps, operaton structures, and protein complexes. We expect that as these tools gets developed and tested, our capability for pathway construction will be significantly improved.

Statistical investigations of the repeatability of nine hybridized comparative DNA genomic microarrays from the 250 gene *Synechococcus* microarrays were studied based upon data provide by Scripts Institute and TIGR. The microarrays consisted of three repeated hybridizations each of the *Synechococcu*s WH8102 printed cDNA microarrays. The hybridizations compared the relative hybridizations of the WH8102 strain with *Synechococcus* strains C129, WH7803, or WH 8113. There were three repeated microarrays for each of the three strains compared with WH8102. Our analysis demonstrated very high within slide repeatability, but often the between slide repeatability was poor. The repeat microarrays run on the same day were more repeatable than those run four months earlier. Some left-right spatial differences of the hybridization were noted on some of the slides. These results were discussed with TIGR personnel, and additional hybridization experiments will be performed to try to identify the source of the lack of repeatability.

Since gene expression microarray expression experiments have not yet been performed on *Synechococcus*, we have been working with microarray data from other collaborations to develop and test some of our ideas presented in the proposal. For example, we have obtained microarray data from 141 Affymetrix chips. Nine of the microarray chips were repeated measurements over time. A statistical examination of the repeat data indicates that the nine repeat microarrays fall into two sets of 4 and 5 microarrays. Within sets, the repeat data are linearly related, but between sets the data do not follow a linear relationship. We have been able to develop an error model for each of the 12,600 genes in the microarrays based on the data from the first four repeat measures. From this model, we can obtain many independent representations of the microarray errors in order to form realistically simulated microarray data that will serve the basis for testing various bioinformatics algorithms applied to the simulated data where the signal will be exactly known and the error will be comparable to that found with real microarrays. Thus, we will have a method to rapidly compare the performance of a number of bioinformatic classification methods.

Collaboration with Others

The ORNL SP3 team has been in close contact with Brian Palenik of UCSD, Tao Jiang of UCR, David Haaland of SNL and his team working towards the goals of the project. Also the team has been interacting with Andrey Gorin, Nagiza Samatova of ORNL/SP2 team and Frank Olkan and LBL on issues relevant to the projects.

David Haaland traveled to TIGR in Rockville, MD to discuss their *Synechococcus* microarray data and experimental procedures. He also discussed with Ian Paulson of TIGR the proposed whole genome *Synechococcus* microarrays that TIGR will generate and hybridize for our GTL project.

Jerilyn Timlin, David Haaland, and Anthony Martino of Sandia met with Arie Shoshani, Frank Olken, and Vijaya Natarajan of LBNL to discuss and implement the new MIAME standards for our future microarray data.

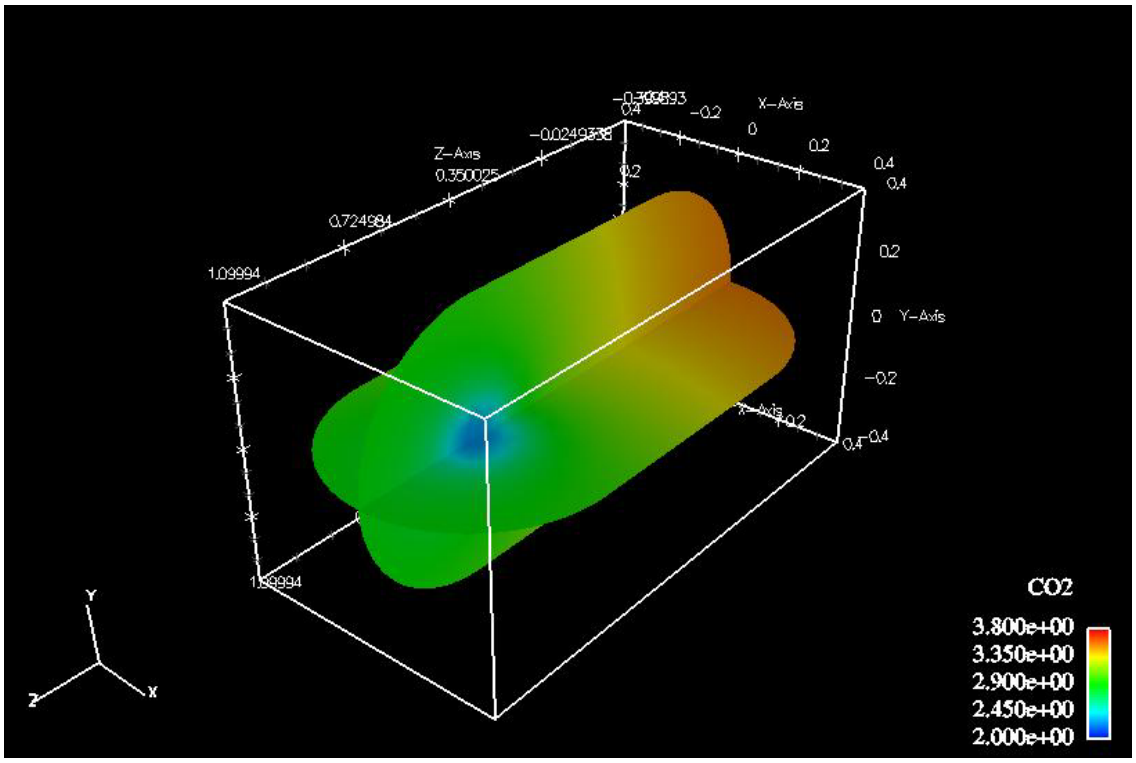
Publications and Presentations

D. Xu, P. Dam, D. Kim, M. Shah, E. Uberbacher, and Ying Xu, "Characterization of Protein Structure and Function at Genome-scale using a Computational Prediction Pipeline", accepted to appear in *Genetic Engineering: Methods and Principles,* Jane Setlow (Ed.), Plenum Press, (by invitation), 2003.

Budget Report (8-02 thru 1-03)

The money allocated to date has been:
SNL: $55K
ORNL: $135K
UCRiverside: $7K

Subproject 4: Systems Biology for *Synechococcus* Sp.



The results of a simple, trial simulation of carbon transport and fixation in a cylindrical representation of an individual *Synechococcus* (shown above) via a suite of reaction-diffusion equations.

This work was funded in part or in full by the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org).

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences.

Introduction

We have begun work on all 4 of the specific aims of SP 4. This includes (arranged by Aim)

1) By comparing the random power-law networks to a real-life system with the same power-law (a yeast 2-hybrid protein network), we have shown that there are considerable topological differences in the two networks. Besides completing the characterization of these networks, we have begun to develop a new methodology to compute protein domain-domain interaction probablilities.
2) We have met with our collaborators from tMSI (the Molecular Sciences Institute) to formulate a general particle model of protein interactions in a cell that includes diffusion and reaction terms as particles that are tracked spatially.
3) We implemented a simple, trial simulation of carbon transport and fixation in a cylindrical representation of an individual *Synechococcus* (shown above) via a suite of reaction-diffusion equations.
4) We have begun the assimilation of expertise in software engineering design principles, software development, scientific domain knowledge, and scientific modeling. We have made good progress towards this goal by implementing a process of dual, simultaneous software engineering and scientific workflows. Progress to date is documented in over 80 pages of work. These efforts place us now in the Design phase of the process and well positioned for delivery of a prototype.

Accomplishments

The main accomplishment related to our first aim, Protein Interaction Network Inference and Analysis, has been to demonstrate that the power law first published by Jeong et al. (Science, 2000, 407, 651-64) is not enough to fully characterize the scale-free nature of protein interaction networks. Preciselly, we find that random networks that follow the same power law as the yeast 2-hybrid protein network have different topological characteristics.

For our second aim, we have been combing the literature (with advice from our experimental collaborators) looking for the currently known parameters that describe the transport and fixation of carbon in *Synechococcus*. We have implemented the data we have found in a simple proof-of-principle simulation using a large system of reaction-diffusion equations to model carbon transport and fixation in *Synechococcus*. We have also worked with collaborators of Brian Palenik to make a realistic geometrical model of *Synechococcus*.

Our work on Aim 3 at this point has been focused on collaborative meetings with our partners to design a software framework to implement the stochastic network modeling as planned.

Within the first six months of the project, we have addressed the major software engineering issues in Aim 4 in problem delineation and constraint identification (see Progress Towards Milestones below). This required attention in both software and scientific development. Accomplishing this has positioned us in the current Design stage with sufficient time for implementation. This type of aggressive forward positioning is important so that we can effectively respond to the modeling application challenges that will arise once the project is generating data.

Progress Towards Milestones

Aim 1: Related to our first milestone, Protein Interaction Network Inference and Analysis, we have implemented and applied several methods to characterize scale-free networks and protein interaction networks (cf. Task 1 in proposal), and we are developing a new methodology to compute protein domain-domain interaction probablilities (cf. Task 2 in proposal).

1. Scale-free and protein interaction networks characterization. Recall that the goal of the first milestone is to infer protein networks by sampling the space of scale-free networks, which is much smaller than the space of random networks. Within that goal, the purpose of the first task to provide insights into the scale-free nature of protein networks beyond the power law that is currently being used. During this quarter we have implemented several network characterization tools. Network automorphism group to detect symmetries, cycle size distribution, degree distribution and extended degree distribution. The degree of a protein is the number of immediate neighbors of that protein in the network, e.g., the number of interactions the protein is involved with. The extended degree of a protein is the degree of the protein, the average degree of the first neighbors, the average degree of the second neighbors, and so on, up to a predefined neighborhood radius. The above characteristics have been tested on four networks. The Yeast 2-Hybrid network, Y2H, published by Ito et al, PNAS, 2001, 98, 4569-4574, this network is composed of 3287 proteins and 4549 interactions. A random network, RN0, composed of the same number of vertices (proteins) and edges (interactions) than Y2H. A random network, RN1, following the same power law than Y2H (i.e., same degree distribution), and a random network, RN2, having the same power law and anti-correlation law for the first neighbor than Y2H (i.e., same extended degree distribution up to the first neighbor).

	Nb. of symmetry classes	Average cycle size	Anti-correlation laws up to 4^{th} neighbor for a protein of degree k
Y2H	1949	37.4	$k,[1/k]^{0.6},[1/k]^{-0.17},[1/k]^{0.4},[1/k]^{0.06}$
RN0	2968	76.5	no correlation found
RN1	2160	28.0	$k,[1/k]^{0.13},[1/k]^{0.16},[1/k]^{0.36},[1/k]^{0.25}$
RN2	1955	40.4	$k,[1/k]^{0.6},[1/k]^{-0.04},[1/k]^{0.27},[1/k]^{0.12}$

The results presented in the above table indicates that RN0, and RN1 differs from Y2H, consequently the power law alone is not enough to fully characterize the Y2H protein network and neighbor's anti-correlation laws have also to be taken into account.

2. Prediction of protein domain-domain interaction probabilities. We are evaluating a new computational technique to predict protein-protein interactions within a cell, using *signature*, a molecular descriptor newly developed (Viso et al. J. Molecular Graphics and Modelling, 2002, 20, 429-438). This descriptor has already been used successfully to predict novel peptide sequences capable of binding to a single target protein, given a dataset of peptides binding to that protein, using a statistical regression technique known as quantitative structure activity relationship (QSAR). Our new technique will use the signature molecular descriptor to predict characteristic *pairs* of protein domain-domain interactions, from an initial data set of interacting proteins, by modifying the standard QSAR to a double-QSAR. If successful, this approach can be used to complement phage-display experiments by 1) making predictions of additional proteins that may interact with the domain being studied, beyond the peptide sequences that came up from the phage-display experiment; and 2) using the characteristic domain-domain signatures to make predictions about binding interactions for proteins beyond the original domain being studied. To apply the signature molecular descriptor to pairs of protein domain-domain interactions we need to start with a data set consisting of interacting domains. In order to develop and test our methodology, we have acquired yeast phage display data from Tong et al paper. The technique will be applied later to phage display data from subproject 1, as it becomes available.

Aim 2: In the area of particle-based cell modeling, discussions have been held with our collaborators at TMSI (Lok, Brent) on two trips to CA, as to how to formulate such a model. Our plan is to create a

general particle model of protein interactions in a cell that includes diffusion and reaction terms as particles are tracked spatially, and that uses the stochastic Moleculizer code already developed at TMSI to generate the reaction equations and rate coefficients needed as inputs. Currently, we are formulating the rules for how reactions will be simulated in the model when there are spatial dependencies, so as to mimic the stochastic model assumption of well-mixed reactants.

Aim 3: Carbon transport and fixation spatial dynamics. The spatial and temporal concentration gradients resulting from the surface transport and internal fixation mechanisms for carbon in *Synechococcus* are of particular interest. We utilize a suite of reaction-diffusion equations to represent such mechanisms and numerically solve these equations with a finite-element solution method.

In order to properly implement this mathematical representation, we require relevant parameters for the transport and fixation of carbon in *Synechococcus* and some understanding of the overall cellular system would be nice as well. Hence, primary activity of late focused on elementary systems study of *Synechococcus*, including research of literature and study of microbiology in general.

Upon discovery of relevant parameters in the literature, we implemented a simple, trial simulation of carbon transport and fixation in a cylindrical representation of an individual *Synechococcus*. The cylindrical representation was based on a slice-plane, confocal microscopy reconstruction of an individual *Synechococcus* cell with the general shape being roughly cylindrical with spherical endcaps. The above image shows this geometrical representation.
The above image also shows the result of said trial simulation, where carbon dioxide is being transported across the surface and a fixation site, or sink, is located along the interior axis of symmetry, anterior side (blue region). Not all parameters (diffusion rate, rates of fixation and activation mechanism, efflux rates, if any) are available as of yet; however, simulation provides opportunity for preliminary investigation of carbon surface transport.

Aim 4: The major first year milestone is a prototypical system for hierarchical modeling. This depends on a series of smaller software development and scientific milestones. We have made substantial progress in meeting that first year milestone as documented in over 80 pages of work in both software and scientific development:

1) Software Development

a) Writing an eight page Software Development Plan entitled "Software development plan for the hierarchical modeling of carbon sequestration in *Synechococcus* sp.";
b) Writing a 25 page Use Case document entitled "Use Cases (or Particular Simulations) for the Hierarchical Simulation Platform" consisting of six Scenarios and work in progress on use case delineation;
c) Writing a nine page Requirements document entitled "Hierarchical Simulation Platform Requirements;" this is a work in progress;
d) Writing a 20 page white paper entitled "Overview of Modeling and Simulation and Object Oriented Software Design Pertinent to Development of a Hierarchical Simulation Platform";

2) Scientific Development

a) Writing a nine page white paper entitled "Current state of the data in the cyanobacterium *Synechococcus*";
b) Writing a 15 page white paper entitled "Population Dynamics, Marine Community Ecology, and Oceanography of Marine Microbes";
c) Writing a white paper entitled "Physiology and Molecular Ecology of *Synechococcus* WH8102;" this is a work in progress.

Collaboration With Others

We currently are working closely with the experimentalists from SP 1 (Palenik, Martino, Lane) on almost all aspects of SP 4 as we begin to gather data and create biologically faithful models. Much of the work we are doing has some initial work that can be leveraged from an earlier project, but our application to *Synechococcus* has required close work with our experimental collaborators.

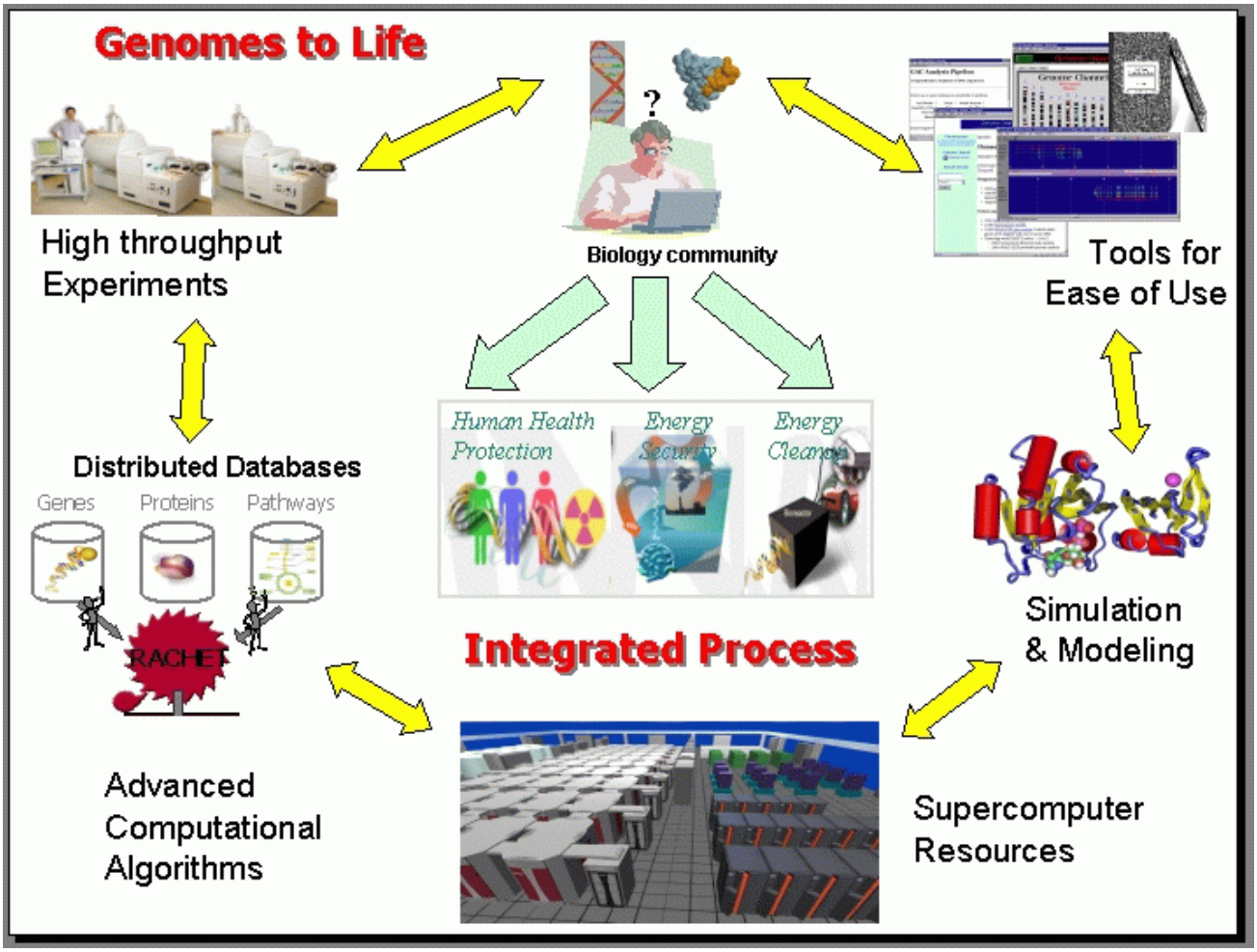
Publications and Presentations

Presentation entitled "GTL October 2002 Report" delivered by Damian Gessler, NCGR, to the first semi-annual Genomes to Life Project meeting in Oak Ridge, TN. on Oct. 31-31, 2002.

Budget Report (8-02 thru 1-03)

The money allocated to date has been:
SNL: $99K
NCGR: Contract is being negotiated
TMSI: Contract negotiations yet to occur
UIUC: $15K

Subproject 5: Computational Biology Work Environments and Infrastructure



This Goal 4 GTL project involves the development of new methods and software tools to help both experimental and computational efforts characterize protein complexes in *Synechococcus*, its regulatory networks, and its community behavior..

This work was funded in part or in full by the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org).

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences.

Introduction

A key to the acceptance of high performance computing in GTL is ease of use and coupling between geographically and organizationally distributed people, data, software, and hardware. Thus an important consideration in the GTL computing infrastructure is how to link the GTL researchers and their desktop systems to the high performance computers and diverse databases in a seamless and transparent way. We propose that this link can be accomplished through work environments that have simple web or desktop based user interfaces on the front-end and tie to large supercomputers and data analysis engines on the back-end.

These work environments have to be more than simple store and query tools. They have be conceptually integrated "knowledge enabling" environments that couple vast amounts of distributed data, advanced informatics methods, experiments, and modeling and simulation.

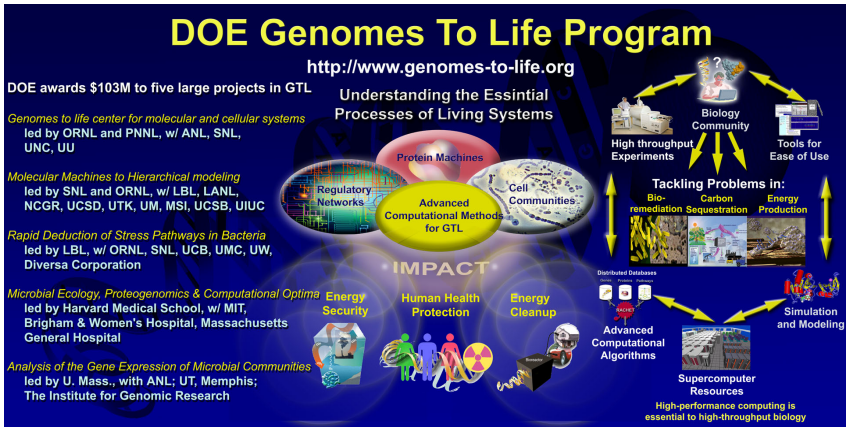
Work environment tools such as the electronic notebooks have already shown their utility in providing timely access to experimental data, discovery resources and interactive teamwork, but much needs to be done to develop integrated methods that allow the researcher to discover relationships and ultimately knowledge of the workings of microbes.

In addition to the development of new computational biology work environments and infrastructure, we plan on supplying the computational resources required by the computational biology methods and algorithms developed in this effort. To this end, arrangements have been made to provide access to ORNL's clusters and supercomputers as well as Sandia's Cplant commodity cluster. We expect that these resources will be significantly employed by the participants, partners, and collaborators in this proposed work.

Accomplishments

The following has been accomplished in the first quarter of this project:

1. Project web site (www.genomes-to-life.org) created and maintained. Including project overview, calendar, job opportunities, publications, and team member pages.
2. Installed a password-protected electronic notebook for the project to facilitate collaboration and data sharing. 30 new pages of material were added this quarter.
3. Supercomputing 2002 conference booth space and posters for DOE Genomes to life initiative and our project in particular were created and set up for the week-long event. Booth was seen by Secretary of Energy Abraham, Undersecretary Card, Dr. Orbach, Jim Decker, and other high ranking DOE managers.



Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences.

4. Began initial design of biology-aware notebook based on the Web services and XML. Identified staff to work on this task once funding is sent from Sandia.
5. Set up a MIAME schema demonstration based on data supplied by Tony. Go to http://sdm.lbl.gov/~opm7/sdmdev/dbs/SGTL_OM/sgtl_query.html to see a sample query page that has been put together for the SGTL database.
6. Began developing the specifications of a general-purpose graph data management system for biological network data. http://www.lbl.gov/~olken/graphdm/graphdm.htm

Progress Towards Milestones

Aim 1. Integrating new methods and tools into an easy to use working environment.
The first of many steps in this process is to develop working environments with transparent access to distributed databases and computational resources. During the reporting period ORNL set up such a work environment for the subproject 3 team to perform their genome-scale protein structure and function predictions.

Aim 2. Develop general-purpose graph-based data management capabilities for biological network data arising from the *Synechococcus* and other studies.
Frank Olken has begun development of a general-purpose graph data management system and has set up a web site to describe the design, features, and potential impact of this tool on analysis of biological networks.

Aim 3. **Develop efficient data organization and processing of microarray databases.** Data provided by Tony Martino with the help of Jerry Timlin of subproject 1 was placed in the SGTL-microarray database, with schema based on the MIAME schema, but has only objects and attributes that Tony found useful. The database also has links to excel spreadsheets, text files, pdf files, and html pages, which are external to the database, just to show this capability as well.
All the web interfaces and data browsing interfaces are generated automatically from the schema by the LBNL-Object-Database-Tools (LODAT) we are using. The underlying system is Oracle.

Aim 4. Develop new cluster analysis algorithms for distributed databases.
Demonstrated new tool for distributed cluster analysis at SC2002 booth. Integration with subproject 2 protein analysis tools will begin in the next quarter.

Aim 5. Establish a biologically-focused computational infrastructure for this effort.
Identified potential hire to work on biology-aware electronic notebook. Began initial notebook design based on web services and XML.

Collaboration With Others

Jerilyn Timlin, David Haaland, and Anthony Martino of Sandia met with Arie Shoshani, Frank Olken, and Vijaya Natarajan of LBNL to discuss and implement the new MIAME standards for our future microarray data from subproject 1.

John Mungler of JICS has been working with Nagiza Samatova in subproject 2 to provide computational resources on ORNL's xtorc cluster for doing protein analysis.

Al Geist of ORNL has been working with Ying Xu in subproject 3 to provide computational resources at ORNL to do his genome-scale protein structure and function predictions on all orfs of *Synechococcus* sp. and two related genomes *Procholorococcus* MIT and MED.

Publications and Presentations

Talk “DOE Genomes to Life Program” presented by Al Geist at the ESnet Steering Committee Meeting. Newport News, VA. March 2002.

Talk “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling” by Al Geist at Supercomputing 2002 conference November 2002.

Budget Report (8-02 thru 1-03)

The money allocated to date has been:
Sandia: $30K
LBNL: $65K
ORNL: $65K (arrived on the last day of this reporting period)
JICS – Contract negotiations yet to occur.

DISTRIBUTION FOR FEBRUARY 2003 GTL QUARTERLY SAND REPORT

5	MS-0885	Grant Heffelfinger, 1802
1	MS-9018	Central Technical Files, 8945-1
2	MS-0899	Technical Library, 9616

Oak Ridge National Laboratories, Sandia National Laboratories, Lawrence Berkeley National Laboratories, Los Alamos National Laboratories, University of California San Diego (Scripps), University of California Riverside, University of Michigan, University of Illinois Urbana/Champaign, National Center for Genome Resources, Molecular Sciences Institute, Joint Institute for Computational Sciences.